A Novel Text Analysis Method: Numerals Reveal the Author

Andrei Zenkov[1,2,*], *Eugene* Zenkov[2], and *Ansgar* Belke[3]

[1]Ural State University of Economics, 620144 Ekaterinburg, Russia
[2]Ural Federal University, 620002 Ekaterinburg, Russia
[3]Universität Duisburg-Essen, 45117 Essen, Germany

Abstract. Two approaches to the statistical analysis of texts are suggested, both based on the study of numerals occurring in literary texts. The first approach is related to the study of the frequency distribution of various leading digits of numerals occurring in the text. This approach is convenient for testing whether a group of texts has common authorship: the latter is dubious if the frequency distributions are sufficiently different. The second approach requires the study of the frequencies of numerals themselves. The approach yields information about the author, stylistic and genre peculiarities of the texts and is suited for advanced study of authorial texts. The hypothesis that I. Ilf and E. Petrov are fake authors of novels "The Twelve Chairs" and "The Little Golden Calf", and they were ghosted by M. Bulgakov, is checked. The frequency distribution of numerals, as well as its cluster analysis, do not confirm this hypothesis.

1 Introduction

The scope of this research pertains to stylometry (statistical study of texts to find individual features of the author's style – in particular, for attribution of texts). The conventional methods – taking into account the frequency of occurrence of certain words and collocations in the text, the average length of words and sentences, etc. [1] – often yield contradictory results, and the very abundance of methods indicates a lack of reliability of each of them individually. Therefore, the emergence of new stylometric techniques is not redundant, and they are all complementary rather than competing.

We have proposed the idea of studying numerals occurring in text as a means of characterizing the author's style [2–5]. The analysis of numerals has many advantages. The results of this analysis allow direct linguistic interpretation (unlike, for example, the neural networks [1] which can successfully recognize the authorship of texts, but the recognition procedure is a black box). The use of numerals in the text is directly related to its authorship, style, and genre (see below).

Our approach to stylometry problems has two varieties. First, we studied the frequency distribution of the leading digits of numerals. The idea may seem bizarre, but it is in line with the Benford's Law [6], according to which in large arrays of numerical data describing various objects and phenomena, numbers starting with digit 1 (their share according to that

[*] Corresponding author: zenkow@mail.ru

law is 30.1 per cent) are more common than those starting with digit 2; the latter are more common than numbers starting with digit 3, and so on. According to our research, the leading digits of numerals in coherent texts are distributed even more unevenly than prescribed by Benford's Law: the proportion of numerals starting with 1 can reach 50 per cent. Usually, the frequency distribution of the leading digits of numerals is characteristic of each author and appears in all (large enough) of his works. Sometimes this allows to check the authorship of texts: if the distributions of the leading digits significantly differ for two texts, then the same authorship of the texts is doubtful.

The 2nd variation of our method consists in analyzing the numerals occurring in text (not their leading digits). The frequency distribution of numerals is also, to a large extent, specific for the author [5]. Each approach has its own advantages and disadvantages. Counting the leading digits makes sense only for leading digits 1, 2, and 3, since the occurrence of subsequent digits is subject to strong fluctuations even in texts by the same author (Fig. 1). Thus, only a small part of the statistical information about the numerals contained in text is available for analysis. In addition, a problem arises with texts in languages in which the numeral *one* is formally indistinguishable from the indefinite article (this is surmountable by switching to an intermediary language without this problem). On the other hand, the information here is presented in a generalized form, which allows to average specific features of individual works of the author.

Analysis of the use of the numerals themselves provides richer information about the author's features of the text and, to a large extent, is devoid of indistinguishability of the numeral *one* and the indefinite article. This article is devoted to the possibilities of both types of our method. We consider a problem related to the Russian literature of the 20th century as well.

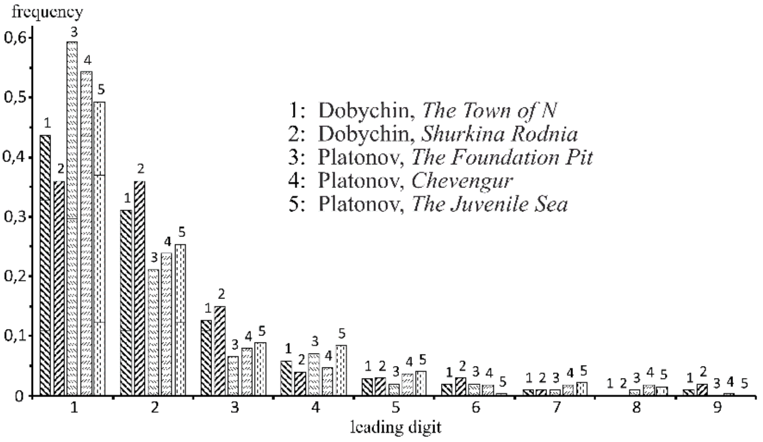

Fig. 1. Frequency distribution of leading digits of numerals in texts by L. Dobychin and A. Platonov.

2 Object and Method of Research

Literary texts by L. Dobychin and A. Platonov are notable for distinct stylistic originality, they have common literary sources and analogues in foreign literature [7]. We will show how this affects the statistics of the use of numerals in their texts.

The literary work of I. Ilf and E. Petrov has repeatedly become the subject of discussion. The novels *The Twelve Chairs* and *The Little Golden Calf* are full of literary allusions; thematically and stylistically they are related to texts by V. Kataev, M. Bulgakov, and others [8]. There is nothing comparable to these two works in the literary heritage of Ilf and Petrov. According to the radical point of view [9], Ilf and Petrov are the fake authors of *The Twelve*

Chairs and *The Little Golden Calf*, and they were ghosted by Bulgakov. In this paper, we will apply our method to a comparative analysis of the corpus of literary texts by Ilf and Petrov. Along the way, we study Kataev's *The Lord of Iron* (1924) and *The Embezzlers* (1926), contemporary to *The Twelve Chairs* (1928) and *The Little Golden Calf* (1931), as well as Bulgakov's *The Master and Margarita*.

We have prepared a computer program that searches in the text for numerals expressed both in numbers and verbally. The texts analyzed were pre-cleaned of numerals that do not reflect the author's creative intent (numbering of pages, chapters, etc.) or accidentally included in idioms (*to the four winds*). Since the analyzed texts have different sizes, correction coefficients were used to equalize the results on the numerals occurrence.

The numerals extracted from the text were displayed on frequency graphs, which directly allowed us to draw conclusions about the author's style. Information about numerals found in texts was also systematized using the hierarchical cluster analysis. The farthest neighbor clustering was used (which exaggerates differences yet provides clearly defined clusters). The smaller the difference in the occurrence of the same numbers in two texts, the greater the similarity (the smaller the "distance" ρ) between the texts, so the Manhattan metric

$$\rho(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{n}|x_i-y_i| \tag{1}$$

was used where x and y are n-dimensional vectors whose components are the absolute frequency of occurrence of the first n natural numbers found in both texts analyzed.

3 Results and Discussion

Fig. 1 shows the frequency distribution of leading digits of numerals in the most voluminous works by Dobychin and Platonov. The leading digits 1, 2 and 3 appear in texts by Dobychin, on the one hand, and Platonov, on the other hand, in very different manner. The visual distinction is supported by Pearson's test [3]. Thus, the analysis of leading digits distribution indicates certain stylistic differences in texts of the two authors.

The results of using an extended statistical method analyzing the occurrence of numerals themselves are richer. Fig. 2 shows the frequency of occurrence of numerals from the range [0, 100] in the same texts by Dobychin and Platonov. Some results:

1. Platonov's texts tend to use numerals more often than Dobychin's texts.

2. Platonov less often resorts to rounding numerals (10, 20, 30, ...), which, together with item 1, may indirectly indicate a greater tendency to detail.

3. the numeral *one* (in various word forms) is the undisputed leader among the numerals found in Platonov's texts. But in Dobychin's texts, the numeral *one* is inferior in frequency to the numeral *two*!

4. Note the understandable rarefaction of the numerals sequence and a decrease in their occurrence as they increase, as well as a noticeable local maximum on the numeral *hundred*, which, of course, plays here the role of an indefinitely large number.

Fig. 3 shows the frequency distribution of numerals in Ilf and Petrov's *The Twelve Chairs* and *The Little Golden Calf*, as well as in Bulgakov's *The Master and Margarita*, and Kataev's *The Embezzlers* and *The Lord of Iron*. For clarity, we restrict the numerals to [1; 50] range. Some conclusions:

1. For all texts, there are peaks in the occurrence of "round" numbers 10, 20, ... , 100, 200, …

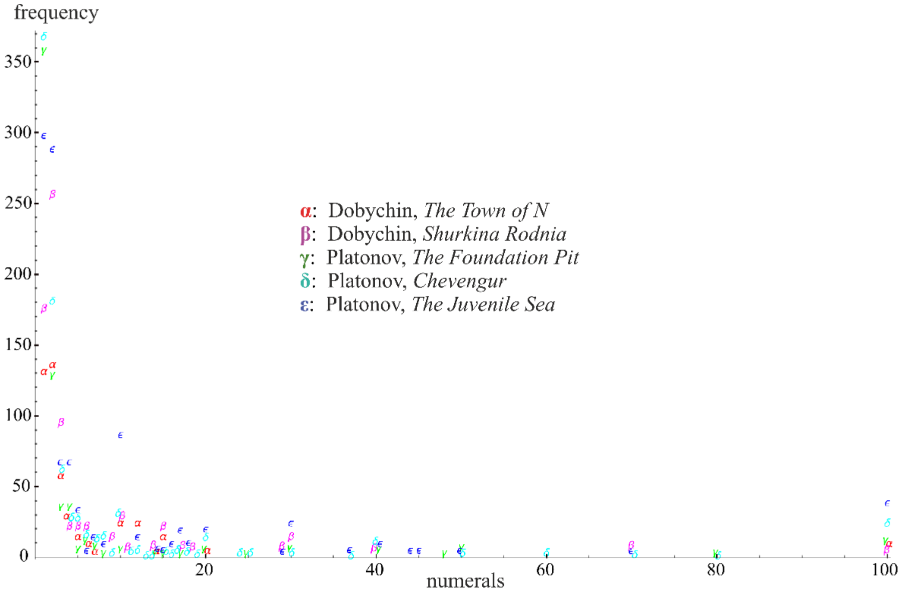

Fig. 2. Frequency distribution of numerals in texts by Dobychin and Platonov.

2. In the texts by Ilf and Petrov, as well as in Bulgakov's *The Master and Margarita*, the numeral 1 has the highest frequency, but in Kataev's texts the number 2 leads.

3. Between *The Twelve Chairs* and *The Little Golden Calf*, there is a conspicuous similarity in the numerals frequency.

4. These two texts are characterized by the greatest variety of numerals.

5. On the contrary, Kataev's texts are distinguished by the least variety of numerals.

6. As for the variety of numerals, *The Master and Margarita* occupy an average position, but the frequencies of the numerals (after the initial frequent *ones* and *twos*) are usually lower than in other texts analyzed. In fact, many numbers occur once.

Based on the frequency distributions of numerals in works by Ilf and Petrov [10], we performed clustering and built a dendrogram (Fig. 4***a***).

All the analyzed texts contain numerals in the range [1; 12]; frequencies of these numerals were used for clustering; n = 12 in formula (1). The numbers to the left of the dendrogram refer to the texts:

1) *The Twelve Chairs*; joint work by Ilf and Petrov, 1927-28; vol. 1 [10],

2) Joint works by Ilf and Petrov, 1932-37, (stories, feuilletons, articles, speeches, vaudevilles, screenplays); vol. 3 [10],

3) *The Little Golden Calf*; a joint work by Ilf and Petrov, 1929-30; vol. 2 [10],

4) Stories, essays, feuilletons by Petrov, 1924-32; vol. 5 [10],

5) Essays, articles, memoirs by Petrov, 1937-42; vol. 5 [10],

6) *One-storied America* (travel essays), 1936, vol. 4 [10],

7) Stories, essays, feuilletons by Ilf, 1923-29, as well as his notebooks from 1925-37; vol. 5 [10].

From the dendrogram, it is clear that the closest (in terms of the numerals use) are ##1 and 2 – joint, mainly literary texts; to this initial cluster, # 3 is soon added with the same characteristic. ##5 and 6 – late non-fiction works – form the next cluster, internally not as unified as cluster {1, 2, 3}. At the last stage, #7 is added – texts by Ilf.

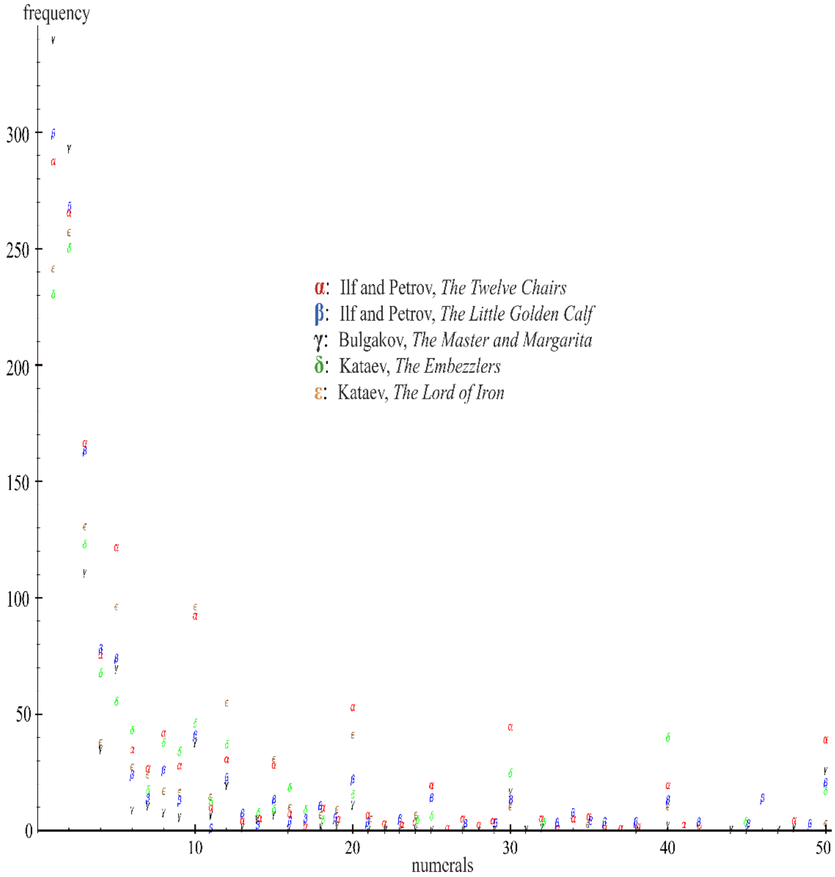

Fig. 3. Numerals in texts by Ilf and Petrov, Bulgakov, and Kataev.

So, from Fig. 4***a*** it follows that the analysis of numerals use can distinguish between *genres* and *authors*.

Fig. 4***b*** shows the results of hierarchical cluster analysis of the numerals occurrence (again from [1; 12] range) in *The Twelve Chairs* (#1) and *The Little Golden Calf* (#2) by Ilf and Petrov, as well as in the works of writers who were named as possible true authors of the two novels – *The Embezzlers* (#3) and *The Lord of Iron* (#4) by Kataev, and Bulgakov's *The Master and Margarita* (#5). Clustering took place in accordance with the generally accepted authorship of texts. The distance between clusters {1, 2} and {3, 4}, not to mention the height of fusion with #5 – is so great that it casts doubt on the hypothesis that Bulgakov or Kataev wrote *The Twelve Chairs* and *The Little Golden Calf*. Of course, taken separately, results of Fig. 4*b* do not confirm the authorship of Ilf and Petrov themselves, but the results of Fig. 4*a* indirectly indicate their authorship. In [11], based on statistics of service words in texts of *The Twelve Chairs* and *The Master and Margarita*, the hypothesis [9] that Bulgakov is the author of *The Twelve Chairs* is also questioned.

So, the analysis of numerals use in texts can be applied to test the texts authorship.

4 Conclusions

The analysis shows that taking into account the occurrence of numerals in literary texts can provide information about the author's, stylistic and genre features of texts. Sometimes, the analysis of numerals occurrence allows to reject the hypothesis of the common authorship of texts.

We believe that our methodology can be a useful addition to the traditional stylometric practices of taking into account the length of sentences and words, the frequency of use of service words and certain significant parts of speech, etc.

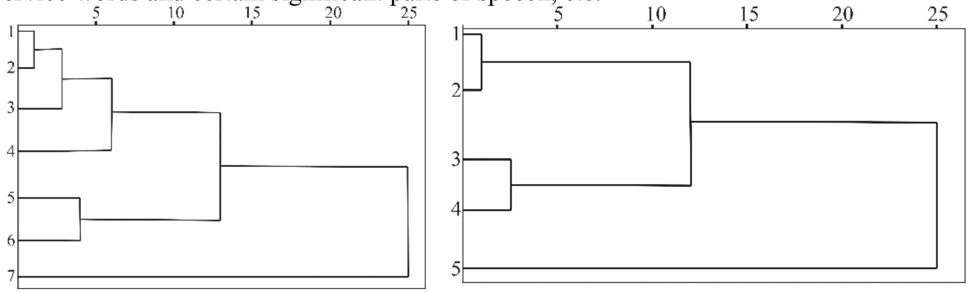

Fig. 4. Results of hierarchical cluster analysis based on the numerals occurrence. ***a***: in texts by Ilf and Petrov. Texts ##1–7, combined into clusters, are indicated in article. ***b***: in texts by Ilf and Petrov (##1, 2), Kataev (##3, 4) and Bulgakov (#5). Texts ##1-5, combined into clusters, are indicated in article. The horizontal scale indicates the "distance" between clusters in conventional units.

Acknowledgements

The work was supported by the grant No. 19-012-00199A from the Russian Foundation for Basic Research.

References

1. N. Tempestt, S. Kalaivani, F. Aneez, ACM Comput. Surv., **50(6)**, 36 (2017)
2. A. V. Zenkov, Computer Research and Modelling, **9** (2017)
3. A. V. Zenkov, J. of Quantitative Linguistics, **25** (2018)
4. A. V. Zenkov, M. Místecký, Glottometrics, **46** (2019)
5. A. V. Zenkov, First Int. Volga Region Conf. on Economics, Humanities and Sports. Paris, Atlantis Press, **114** (2019)
6. F. Benford, Proc. of Amer. Philos. Soc., **78** (1938)
7. V. V. Eidinova, Andrei Platonov's "Land of Philosophers": Problems of Creativity. Issue 5: Based on the materials of the Int. Sci. Conf. dedicated to the 50th anniversary of A.P. Platonov's death (2003)
8. Yu. K. Ščeglov, *The Novels by Ilf and Petrov. Readers's Companion* (2009)
9. I. Amlinski, *12 Chairs from Mikhail Bulgakov* (2013)
10. I. Ilf, E. Petrov, *Collected works in 5 volumes* (1961)
11. L. B. Mironova, O. V. Vereshchagina, Baltic Humanitarian J., **8** (2019)